SCIENCE CHINA
Life Sciences

• RESEARCH PAPERS •

December 2011 Vol.54 No.12: 1129–1133
doi: 10.1007/s11427-011-4256-9

De novo transcriptome assembly of RNA-Seq reads with different strategies

CHEN Geng[1], YIN KangPing[1], WANG Charles[2] & SHI TieLiu[1*]

[1]*Center for Bioinformatics and Computational Biology, Institute of Biomedical Sciences, School of Life Science, East China Normal University, Shanghai 200241, China;*

[2]*Functional Genomics Core, Beckman Research Institute, City of Hope Comprehensive Cancer Center, Duarte, CA 91010, USA*

Received October 6, 2011; accepted November 15, 2011

De novo transcriptome assembly is an important approach in RNA-Seq data analysis and it can help us to reconstruct the transcriptome and investigate gene expression profiles without reference genome sequences. We carried out transcriptome assemblies with two RNA-Seq datasets generated from human brain and cell line, respectively. We then determined an efficient way to yield an optimal overall assembly using three different strategies. We first assembled brain and cell line transcriptome using a single *k*-mer length. Next we tested a range of values of *k*-mer length and coverage cutoff in assembling. Lastly, we combined the assembled contigs from a range of *k* values to generate a final assembly. By comparing these assembly results, we found that using only one *k*-mer value for assembly is not enough to generate good assembly results, but combining the contigs from different *k*-mer values could yield longer contigs and greatly improve the overall assembly.

RNA-Seq, *de novo* transcriptome assembly, next generation sequencing

Citation: Chen G, Yin K P, Wang C, *et al*. *De novo* transcriptome assembly of RNA-Seq reads with different strategies. Sci China Life Sci, 2011, 54: 1129−1133, doi: 10.1007/s11427-011-4256-9

The output of the next-generation sequencers is increasing dramatically while the cost is sharply decreasing. RNA-Seq technology enables us to investigate the transcriptome more comprehensively than microarrays and is becoming more popular for various gene expression studies [1–9]. During the data analysis, the transcriptome sequencing reads are usually first mapped to the reference genome sequences or transcriptome databases if they are available. However, the genomes of most species still have not been sequenced and high-quality-assembled reference genomes are lacking for most organisms. Therefore, *de novo* transcriptome assembly becomes the first analysis step for those unsequenced organisms. Furthermore, *de novo* transcriptome assembly can help researchers further investigate the genes that are missing from the reference genomes due to the incompleteness of reference sequences for those sequenced organisms [10]. Accordingly, *de novo* transcriptome assembly is an important approach for carrying out transcriptomics studies.

Up to now, several pieces of software used for *de novo* assemble transcriptome based on short RNA-Seq reads have been developed. Unlike the overlap-layout-consensus approach from the Sanger Sequencing Method, which is widely implemented in the assembly algorithms for the long reads, the next-generation sequencing assembly programs mainly use the de Bruijn graph approaches. This approach can effectively handle huge amount of short sequencing reads. Velvet [11], ABySS [12], Trans-ABySS [13] and Trinity [14] all use the de Bruijn graph algorithm to process short reads and assemble those related reads into contigs or scaffolds. On account of that transcriptome coverage levels depend highly on the gene expression levels and variations in isoforms, gene families and the repetitive sequences

*Corresponding author (email: tlshi@sibs.ac.cn)

cause many instances of ambiguities, the contiguity of transcriptome assembly is rarely high.

For those organisms that have no high-quality-assembled reference genomes, it is urgent to find an effective *de novo* transcriptome assembly approach to gain their transcript sequences for further inference and annotations of their gene/isoform structures and potential functions. Moreover, further analyses on the assembled contigs are directly dependent on the completeness of the reconstructed transcript sequences from the transcriptome sequencing sample. If the *de novo* transcriptome assembly process yields a large number of short contigs, it is difficult to analyze these short contigs and generate meaningful results. By contrast, the optimal assembly strategy could generate longer and more complete transcript sequences, which make the analysis steps much easier and increase the likelihood of a more meaningful research results. Consequently, it is essential to choose an effective *de novo* assembly strategy to reconstruct the transcriptome as complete as possible.

Those de Bruijn-based assemblers usually have two important parameters: *k*-mer length and the value of coverage cutoff. The length of *k*-mer determines the number of such *k*-mers for a read to be divided. Coverage cutoff is mainly used to remove the artifacts caused by sequencing errors or variants. Since these two parameters largely determine the performance of those de Bruijn-based assemblers, using different values of *k*-mer and coverage cutoff in the assemblies will generate different assembly results [15]. Generally speaking, a single value for *k* and coverage cutoff is unlikely to generate the optimal assembly results. To assess how the values of *k*-mer length and coverage cutoff influence the overall assembly results, we tried three different assembly approaches to assemble two transcriptome sequencing datasets from human brain tissues and 10 mixed cell lines (see Materials and methods). We first assembled these two datasets using a single *k*-mer value, and then we tried a series of *k*-mer length and coverage cutoff values with Velvet assembler. We also tested the strategy that combines the assembled contigs from different *k*-mer values to yield a final assembly. Lastly, we compared the assembly results from these three different strategies and determined the optimal one.

1 Materials and methods

1.1 Data usage

In this study, we used two transcriptome sequencing datasets from two human reference RNA samples established by the MicroArray Quality Control (MAQC) project [16] with standard Illumina next-generation sequencing technology: the Universal Human Reference RNA (UHRR) from 10 human cell lines of various origins [17] and the Human Brain Reference RNA (HBRR) from several regions of the brain of 23 adult donors. These two datasets consisted of ~59.46 million and ~53.24 million of single-end sequencing reads for cell lines and brain, respectively. The sequencing reads are 35 bp in length. These two datasets used in this study can be downloaded from NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession number GSE30222. The human reference genome sequences of hg19 were downloaded from UCSC (http://genome.ucsc.edu/).

1.2 Software usage

We used the Velvet [11] package (version 1.1.06) to conduct the brain and cell line *de novo* transcriptome assemblies. Velvet is available at http://www.ebi.ac.uk/~zerbino/velvet/. To test different assembly strategies, we also used two contributed programs in the Velvet package: VelvetOptimiser and AssemblyAssembler. VelvetOptimiser (developed by Simon Gladman) is a wrapper script designed to assist Velvet to optimize the assembly. AssemblyAssembler is designed by Jacob Crawford to automate a directed series of assemblies using the Velvet assembler. After finishing the assembling, the assembled contigs from three different assembly strategies were aligned to the human reference genome hg19 using Blat (version 34) [18] with -trimT option enabled. We then used the criteria of 90% identity and 90% coverage to remove those contigs that cannot be aligned to the human reference genome.

2 Results

2.1 *De novo* transcriptome assembly

The goal of our study is to investigate how to improve *de novo* transcriptome assembly results with an effective strategy that generate an optimal overall assembly, using Velvet assembler. Good assembly approaches can optimize the overall assembly results and reduce the redundancy of short contigs, and generate longer and more complete contigs. This is vital for further analyses that rely on the *de novo* assembled transcript contigs on account of that longer contigs can be mapped and analyzed more easily compared with the redundant shorter contigs. Our testing datasets were two transcriptome sequencing data of human brain tissues and 10 mixed cell lines from the MAQC [16] project.

To determine the combination of the parameters for an optimal overall assembly, we tried three different strategies. First, we assembled the brain and cell line transcriptome sequencing datasets using Velvet with *k*-mer length of 21 and other parameters as their default values. Second, we searched the *k*-mer values ranging from 17 to 33 for the optimum, estimating the expected coverage, and then searched for the optimum coverage cutoff using the contributed software VelvetOptimiser in the Velvet package. Third, we conducted Velvet assemblies using default pa-

rameter values across 17 to 33 of *k*-mer length and took the contigs from these assemblies as input for a final assembly by AssemblyAssembler on two datasets of the brain and cell line. After completing these assemblies, we compared their assembly results to determine the best strategy.

2.2 Comparison of assembly results

For each assembly strategy, only those contigs longer than 100 nucleotides were kept and shorter contigs were removed. We found that for both brain and cell lines, the first assembly strategy generated the largest number of contigs but many of them are short in length (brain: range from 100 bp to 3336 bp; cell lines: range from 100 bp to 3182 bp) compared with the other two strategies (Table 1). For the second assembly strategy, we used VelvetOptimiser to search the *k*-mer length range from 17 to 33 for the optimum assembly. VelvetOptimiser found the best assembly results are with *k*-mer length of 25 and coverage cutoff of 2.83 for both brain (range from 100 bp to 8474 bp) and cell lines (range from 100 bp to 6087 bp). The third approach generated the least number of contigs (79515 for brain and 79367 for cell lines), but overall its assembled contigs are longer than those from the other two strategies (brain: range from 100 bp to 12895 bp; cell lines: range from 100 bp to 9534 bp).

We further compared the overall assembly results from these three different strategies. For brain transcriptome, the N50 contig length of these three approaches are 244, 300 and 479 bp; and for cell line transcriptome, N50 are 249, 303 and 513 bp, respectively (Table 1). The maximum contig length from the third strategy is the longest (12895 bp for brain and 9534 bp for cell lines), while the first strategy generates the shortest maximum lengths (3336 bp for brain and 3182 bp for cell lines). However, the first strategy obtains the largest total contig length (25.2 Mb for brain and 26 Mb for cell lines) and the second strategy yielded the least total contig length (18.4 Mb for brain and 19.6 Mb for cell lines). Although the sum of contig lengths generated with the first approach is the largest, those contigs might contain more redundancies. The contig length distributions for these three strategies are shown in Figure 1. Although the first strategy generates the largest number of assembled contigs compared to the other two approaches for brain and cell lines, these contigs generated from the first strategy are shorter than the contigs from the other two strategies. In general, the first and the second assembly strategy yielded short and discontinuous contigs, while the third approach increased the continuity of contigs and generated longer contigs.

2.3 Mapping assembled contigs onto the human genome

To select all possible correctly assembled contigs from the whole generated contigs we mapped all the assembled con-

Table 1 Comparison of assembly results for three different strategies[a)]

Contig statistics	Brain			Cell lines		
	First strategy	Second strategy	Third strategy	First strategy	Second strategy	Third strategy
Number	114715	73122	79515	117624	78082	79367
Min (bp)	100	100	100	100	100	100
Max (bp)	3336	8474	12895	3182	6087	9534
N50 (bp)	244	300	479	249	303	513
Mean (bp)	219.8	251.9	328.3	221.4	251.2	339.1
Median (bp)	163	164	184	163	161	188
Total length (Mb)	25.2	18.4	26.1	26	19.6	26.9

a) The first strategy sets *k*=21 for Velvet to carry out assembly; the second strategy represents using VelvetOptimiser to search *k* values from 17 to 33 to find an optimal assembly; the third strategy uses AssemblyAssembler to combine the assembled contigs of *k* from 17 to 33 to yield a final assembly.

Table 2 Aligned assembled contigs for three different approaches[a)]

Aligned contig statistics	Brain			Cell lines		
	First strategy	Second strategy	Third strategy	First strategy	Second strategy	Third strategy
Number	109920	70367	75330	112662	74976	75435
Min (bp)	100	100	100	100	100	100
Max (bp)	3336	8474	12895	3182	6087	9534
N50 (bp)	250	308	494	254	312	526
Mean (bp)	223	256.1	335.1	224.8	255.7	345.9
Median (bp)	166	166	188	165	164	192
Total length (Mb)	24.5	18	25.2	25.3	19.2	26.1

a) The three strategies are the same as those in Table 1. The assembled transcript contigs from these three approaches were aligned to the human reference genome hg19 using 90% identity and 90% coverage as threshold.

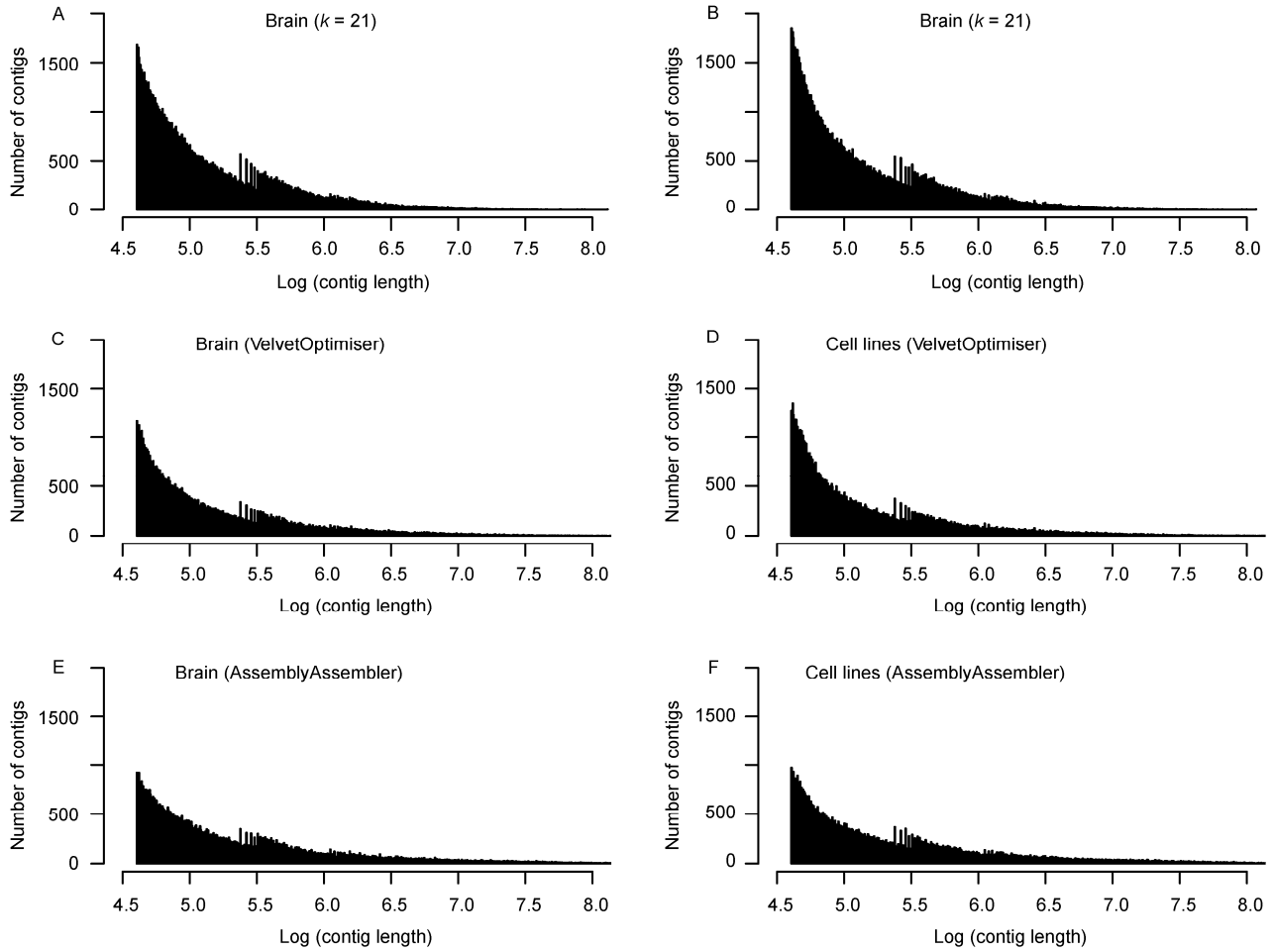

Figure 1 Contig length distributions of three different strategies for brain and cell lines. A and B, The *k*-mer length was set as 21 for Velvet and other parameters were used as default. C and D, VelvetOptimiser was used to search *k* values from 17 to 33 to find an optimal assembly. E and F, AssemblyAssembler was used to combine the assembled contigs with *k*-mer length from 17 to 33 to generate a final assembly. The distribution of contig lengths for the *x*-axis is presented in log scale.

tigs from these three different strategies to the human reference genome hg19 using Blat [18]. Considering that the sequences of the transcribed RNAs might be changed by RNA editing or contain variations, and the human genome sequences are very complex and might make it difficult for the aligner to match all those *bona fide* transcript contigs to the reference sequences, we finally chose 90% identity and 90% coverage as the threshold. With these criteria, 4795 brain contigs (range from 100 bp to 1046 bp) and 4962 cell line contigs (range from 100 bp to 1088 bp) were removed from the assembled contigs of the first strategy; 2755 brain contigs (range from 100 bp to 1400 bp) and 3106 cell line contigs (range from 100 bp to 962 bp) were removed from the second strategy; 4185 brain contigs (range from 100 bp to 4916 bp) and 3932 cell line contigs (range from 100 bp to 4491 bp) were removed from the third strategy. Those discarded contigs might result from the wrong assemblies or the expressed novel transcript contigs that are missing from the human reference genome due to the incompleteness of the human reference genome [10].

After removing the contigs that cannot be mapped to hg19, we compared the aligned contigs from the three approaches. For the first assembly strategy, 109920 brain contigs (range from 100 bp to 3336 bp) and 112662 cell line contigs (range from 100 bp to 3182 bp) were aligned. For the second assembly strategy, 70367 brain contigs (range from 100 bp to 8474 bp) and 74976 cell line contigs (range from 100 bp to 6087 bp) were mapped. For the third assembly strategy, 75330 brain contigs (range from 100 bp to 12895 bp) and 75435 cell line contigs (range from 100 bp to 9534 bp) were aligned. The detailed statistics of those aligned contigs for the three approaches can be found in Table 2. Comparing the results, we found that the third strategy achieved the best assembly results, while the second approach is superior to the first one. Therefore, the most efficient way to improve the *de novo* transcriptome is to combine the assembled contigs from different *k*-mer lengths together to generate a final assembly.

3 Discussion

We conducted *de novo* transcriptome assemblies on two human RNA-Seq datasets generated from the brain and cell line using Velvet with three different approaches. We found that the best way to achieve an optimal overall assembly is to combine the assembled contigs from different *k*-mer length to produce a final assembly. The length of *k*-mer and the value of coverage cutoff are the two most important parameters for Velvet in assembling. Only one *k*-mer length for the assembly is not enough to obtain a good overall assembly. To figure out that which *k*-mer length and coverage cutoff can generate the optimal results, a range of values for these two parameters have been tested. Furthermore, our results also demonstrate that the contig sets from those assemblies of different *k*-mer values have overlaps among them and they can be combined together to improve the assembly results and yield longer contigs.

Transcriptome assembly programs that have been developed currently can mainly be divided into two categories, based on the requirement for a reference genome sequences (genome-guided) or not (genome-independent or *de novo*). The genome-guided methods (such as Cufflinks [19] and Scripture [20]) map the short reads onto the reference genome sequences and assemble the mapped reads into transcript fragments using the mapping information. The other category is *de novo* assemblers that are based on de Bruijn graphs and does not need the reference genome sequences. Although the genome-guided methods can generate the best assembly results in principle, sequencing genome in the past was costly and time-consuming. These factors prevented the sequencing technologies from being widely used other than on a few model species. Moreover, considering the complexities of the genome and transcriptome, it is difficult to obtain a complete and fully annotate reference genome. Consequently, *de novo* transcriptome assembly approach is very useful in RNA-Seq data analysis for various organisms.

In the future, if different lengths of RNA molecules can be fully sequenced in one run, there will be no need for transcriptome reconstruction and data analysis will become much easier. However, it still has a long way to go to reach this goal [21]. Our study can provide guidance for those researchers that need to carry out *de novo* transcriptome assemblies and help them to choose an efficient assembly approach to improve their *de novo* assembly results.

This work was supported by the National Basic Research Program of China (Grant Nos. 2010CB945401, 2007CB108800) and National Natural Science Foundation of China (Grant Nos. 30870575, 31071162, 31000590), and the Science and Technology Commission of Shanghai Municipality (Grant No. 11DZ2260300).

1 Marioni J C, Mason C E, Mane S M, *et al*. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res, 2008, 18: 1509–1517
2 Sultan M, Schulz M H, Richard H, *et al*. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 2008, 321: 956–960
3 Mortazavi A, Williams B A, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods, 2008, 5: 621–628
4 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009, 10: 57–63
5 Maher C A, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. Nature, 2009, 458: 97–101
6 Chepelev I, Wei G, Tang Q, *et al*. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Res, 2009, 37: e106
7 Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol, 2010, Chapter 4: Unit 4.11. 1–13
8 Pflueger D, Terry S, Sboner A, *et al*. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res, 2011, 21: 56–67
9 Chen G, Yin K, Shi L, *et al*. Comparative analysis of human protein-coding and noncoding RNAs between brain and 10 mixed cell lines by RNA-Seq. PLoS ONE, 2011, 6: e28318
10 Chen G, Li R, Shi L, *et al*. Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. BMC Genomics, 2011, 12: 590
11 Zerbino D R, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res, 2008, 18: 821–829
12 Birol I, Jackman S D, Nielsen C B, *et al*. De novo transcriptome assembly with ABySS. Bioinformatics, 2009, 25: 2872–2877
13 Robertson G, Schein J, Chiu R, *et al*. De novo assembly and analysis of RNA-seq data. Nat Methods, 2010, 7: 909–912
14 Grabherr M G, Haas B J, Yassour M, *et al*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol, 2011, 29: 644–652
15 Chitsaz H, Yee-Greenbaum J L, Tesler G, *et al*. Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets. Nat Biotechnol, 2011, 29: 915–921
16 Shi L, Reid L H, Jones W D, *et al*. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol, 2006, 24: 1151–1161
17 Novoradovskaya N, Whitfield M L, Basehore L S, *et al*. Universal Reference RNA as a standard for microarray experiments. BMC Genomics, 2004, 5: 20
18 Kent W J. BLAT—the BLAST-like alignment tool. Genome Res, 2002, 12: 656–664
19 Trapnell C, Williams B A, Pertea G, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28: 511–515
20 Guttman M, Garber M, Levin J Z, *et al*. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol, 2010, 28: 503–10
21 Zhou X, Ren L, Li Y, *et al*. The next-generation sequencing technology: a technology review and future perspective. Sci China Life Sci, 2010, 53: 44–57